Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 119-125

Research Article

Time series modelling and forecasting for predicting Covid19 Case Load using LSTM

[1]Sellam.V, [2]Mohit Gorakhapuriya, [3]Avani Mishra, [4]Prince Kevadiya

[1]Assistant Professor, Department of Computer Science and Engineering
[2,3,4]Student,SRM Institute of Science and Technology, Ramapuram Campus, Chennai

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

ABSTRACT- The epidemic of the Novel Coronavirus across the globe has influenced the globe overall and caused a large number of death results. This remains as an unfavorable admonition to general wellbeing and will be set apart as perhaps the inordinate pandemics in the history. Inorder to validate and analyse, the details was taken from COVID-19. The detail contains daily tallies of confirmed, relieved and demise cases. Likewise, it includes extra data with respect to patients testing present in various states and the outcomes isolated in confirmed and invalidated cases. With the data provided it allows infected person to get the proper treatment and timely quarantine. The proposed paper utilizes Long Short Term Memory (LSTM) networks for sequential prediction of data. The networks are viable apparatuses in short-term time series gauge the COVID-19 confirmed cases. It is a complex gated memory unit made to disappearing gradient issues restricting the effectiveness of a basic Recurrent Neural Network (RNN). Here Neural Network is used to solve the complex operation on the dataset. The result demonstrate that the LSTM Network is executed with various activation functions by utilizing a exponential linear unit brought about better execution for determining the complete number of COVID-19 cases. With the timely observations the corona virus state can be effectively monitored and the proper treatment can be assigned for the infected ones.

KEYWORDS: Long Short Term Memory (LSTM), COVID (Corona Virus Disease), Recurrent Neural Network (RNN), Least Absolute Shrinkage and Selection Operator (LASSO).

I.INTRODUCTION

Precise and quick analysis of COVID-19 suspected cases assumes significant part in timely isolation and appropriate treatment. Building up a deep learning-based model for COVID-19 diagnosis on CT is useful to detect the effects of SARS-CoV-2. A feeble supervised deep learning framework was created utilizing 3D CT volumes for COVID19 variations.[17] During the assessment with anticipating of a time series, it is proposed to plot the time series information and separate for unique highlights. This might appear to be a monotonous methodology however gives legitimate experiences on the spread and picks the modeling road map. Consequently , increment of the quantity of death information as input, gives defects in anticipating, contrasted with utilizing just the quantity of total cases provided.[13] However, Damped Trend offers higher outcomes to LSTM Networks in anticipating whole range of COVID-19 cases.

The pre-observation on the patterns of the Covid-19 transmission and the LSTM networks are efficient methods in the short term time series prediction of the covid confirmed cases are hereby defined. The data is pictured which gives a significant set-up of tools and methods for acquiring a subjective agreement followed by the element choice and the expectation. The network involves a LSTM Block which is arranged as a cycle for recursive multi-step time series forecasting values of various future days.[11] It plans a tendency to rent 3 LSTM fashions for short-run statement the unfold of COVID-infections amongst designated states in India. Indian states were picked with COVID-19 hotpots in phrases of the fee of infections and evaluate with states where infections have been contained or reached their peak.

This paper is a proposal to try each univariate and multivariate time collection prediction methods and examine their overall performance for temporary (4 days ahead) forecasting, with a two months beforehand forecast the use of chosen LSTM models.[19] The existing visualization and analysis of the COVID-19 infections and furnish open supply software that can be used as greater records receives accessible and additionally utilized to different nations and regions.[2]

II. BACKGROUND AND RELATED WORK

For forecasting the usage of time collection analysis, a number of models, such as LSTM networks and Auto-regressive Integrated Moving Average (ARIMA) mode, are involved. Some primer exploration for COVID-19 time sequence forecasting the use of ARIMA have furthermore been finished. In any case, two roads of query – estimating for over one days later and the evaluation of various LSTM styles with ARIMA model – are yet to be comprehensively explored which learn about exploresthe overall performance of various LSTM fashions.[7]

Moreover, given that it is profitable to consider the overall act of these time-string forecasting ways for anticipating a couple of values in the future alternatively than the single estimate, the concept of k-period overall performance metrics is introduced.[6] These k-period overall performance metrics lengthen current performance metrics used to calculate a single prediction to the common case of calculating more than one forecast in the future. While the proposed approach of k-period overall performance metrics can be used to prolong any performance metric, we mainly pick used two methods one was Mean Absolute Percentage Error (MAPE) and another was Median Symmetric Accuracy (MSA) for acquiring the k-period overall performance metrics used in this study.

The important benefactions of this paper are: The study and evaluation of the overall representation of infinite LSTM architectures vis-à-vis the ARIMA model in time series evaluation and forecasting of country-wise COVID-19 cases for four countries.[10] It is recommended and proposed k-period overall performance metrics for estimating and calculating the performance of time-gathering forecasting algorithms the point forecasts are made for some of the periods in the future. While this explanation can be used for any overall performance metric in time sequence forecasting, and expand the definition of MAPE and MdSA metrics to outline k-period kMAPE (kMAPE) and k-period MdSA (kMdSA), respectively, in particular.

Methods focused on searching parameterized tuning to be a black-box optimization issue , and use basic methods to find parameter spaces.[5] A most modern related find out about the inquiry-based approach is BestConfig, which uses the division and-wander sampling strategy and the recursive bound-and-search consideration to exactly tune classifications with limited support for mainstream frameworks. Suggested a commotion inclination calculation, alluded to as synchronous bother stochastic estimation (SPSA), to enhance Hadoop's presentation.

The PA-based methodology takes in general attributes the utilization of a fine-grained estimate of the run-time nation of the program, and offers a test method to reenact the task performance way and anticipate performance.[11] Overall performance anticipation mannequin formed means of technique additionally known as a cost-based model. MRTuner proposes a PTC life sized model to consider the expense of equal execution between unmistakable obligations and plans a climate well disposed inquiry calculation to select the best quality level performance plan. RFHOC offers to partition the guide and limit periods of the Hadoop occupations two or three essential tasks, and afterward a Random Forest is utilized to research the worth of each major activities.

The ML-based technique educates the overall presentation forecast mannequin use of coaching facts accumulated through the amassing model. [3] The peering out life-sized model would then be able to find the best settings dependent on the expectation results from the forecast model, which is similar to our task. [4] For instance, A DAC was made concerning Hierarchical Modeling (HM) and a GA. HM is an expectation life-sized model developed the utilization of relapse trees, and GA is accountable for paying special mind to most productive designs in boundary space.[18] ALOJA-ML makes use of more than one desktop studying fashions to predict Hadoop application presentation. CBM developed through a registering gadget becoming more acquainted with calculation, and utilize Latin hypercube testing (LHS) to create and look for most helpful boundary arrangements of apportioned message framework built a Spark application generally execution forecast life-sized model through irregular backwoods and find the best setups the use of a hereditary calculation advocate an SVM based absolutely procedure to precisely tune the designs of Hadoop programs. Wang et al. built a Spark programming by and large execution expectation life-sized model by utilizing paired grouping and multi-arrangement.[18] During the ML-based methodology, we exclusively require to analyze the design and execution season of the responsibility, overlooking the significant places of the inward walking measures. Along these lines, this methodology can be utilized for boundary tuning for endless structures. Be that as it may, to accomplish a right in general exhibition expectation model, a decent measured amount of schooling records wishes to be accumulated to train the model, which is quite tedious. ATCS varies by accomplishing a high-exactness generally speaking execution forecast life-sized model including a little quantity of training information. ATCS permits moving across extraordinary gatherings.

III. PROPOSED METHODOLOGY

Due to quick rise in covid cases proposed system allow affected person to give medical treatment on critical time and quarantine them. Making such deep learning model allow automatic diagnosis of covid-19 effect and stop widespread of virus. The framework was made using CT scan volumes for corona virus classification. The proposed system at the time of analysis is suggested to plot the data and analysis for specific attribute. It may seem a slow motion but provides perfect visual image on the distributed and helps choose the representation road map.[9] Consequently, more number of data of death give more accurate result with lower error, in compare to only number of case in input.

Proposed method can be better than existing approach as to improve performance developer provide high durable parameter while in proposed method it is simple to implement. It does not have the scalability problem.[12] This approach supports less processing and storage resource. It is quick and efficient with precise as the state-of-the-art algorithm. Proposed method cost low computational cost.

Here neural network is used to solve complex operation on dataset. Neural network is architect of connected neuron. It is combination of more than one algorithm. Recurrent neural network (RNN) is category of neural network used to deal with temporary information. The neuron in this method have cell/state memory with input which is operated on according to internal state. There is repeating layer in RNN which store data for short duration which is perfect fit for LSTM.

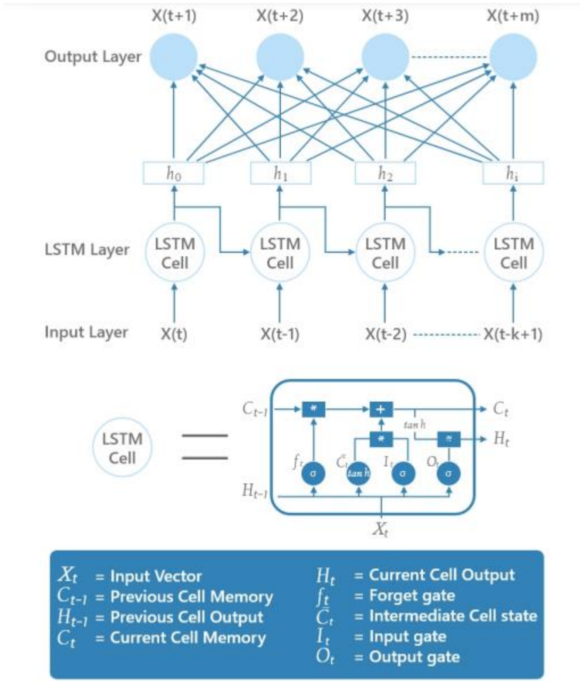


Fig 1. Long short-term memory(LSTM)

Long short-term memory (LSTM) is Recurrent neural network (RNN) network which predict durable pattern in data. It is achieved by four layer of recurring module which is communicating with each other. LSTM is a complex gated memory unit made to disappearing gradient issues restricting the effectiveness of a basic RNN [19]. Because of significant time, the inclination turns out to be excessively little or enormous, which brings about disappearing gradient issue. This problem shows up during the preparation, where backpropagates and makes the system run, while the loads nearly do not change. Basically, LSTM has three entryways controlling the data stream termed input, forge, and output gates, these gates are framed just with calculated elements of weighted aggregates; the loads can be acquired during preparing by backpropagation.[6] The cell state is overseen by means of the information input gate and the forget gate. The O/P is produced from the output gate, which addresses the memory coordinated for use. This system allows the network retaining for quite a while which is missed the conventional RNNS. LSTM networks are widely utilized in DL.[6] LSTMs have been made to technique sequences of statistics and multiplied on common RNN through using cells which store data in reminiscence for long pattern, and a set of gates to manipulate the flow of this reminiscence info.

IV. Modules Description

A. List of Modules

1. Data Visualization

Data visualization is a graphical representation of the data so it easy to get overview of the large dataset. Different type of graph is used to represent data in for of bar, line chart, pie chart etc. To communicate meaning of the data it is quick way to represent pattern behind data. This practice can help find factors affecting behavior of the data. In this module the data set is represent in form of the graph to get idea of which data or field should be reduced to prepare data for algorithm. Numerous time it is repeated and dataset iss reduced on basis of the graph pattern.

Data Visualization is a big step in reducing and cleaning data and prepare for further step. It is used to se field having partial data by providing specific parameter in graph. Comparison of data shown using graph give idea of data integrity and consistency.

2. Feature Selection

Feature selection is the process in machine learning to improve the performance of the algorithm. The data which is used in model have huge impact on performance during training of the model. It is the selection of feature automatically or manual to contribute most to the prediction of the result. To select feature it should be in such a way that it reduce the opportunity to make prediction on basis of irrelevant data, give more accuracy, reduce time taken for training the model.

In this module parameter of the model is selected to minimize the noise effect while predicting the result. To remove unnecessary data field various factor are considered to make training data more fitting for predictive model. It involves splitting of data into 20% training data and 80% data for testing.

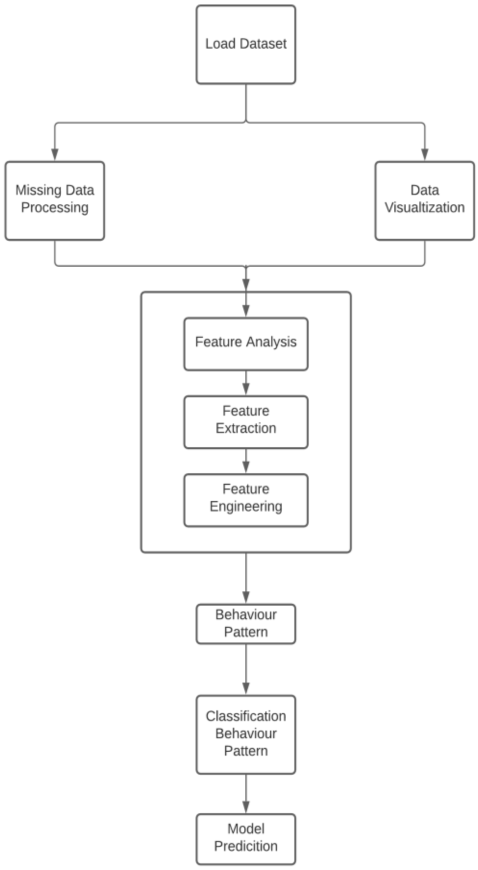
3. Prediction

In this module prediction algorithm is used on the data which is cleaned and prepared for model. Long short-term memory (LSTM) algorithm is used here so result can be predicted on bases of short term retained information using Recurrent neural network (RNN). In this module algorithm are used having three-layer as input gate, hidden gate, output gate. In here sequential algorithm is used to implement the model. LSTM algorithm is defined with sequential algorithm and three-layer network of input layer, hidden layer, output layer. It uses optimizer Adam with loss as binary cross entropy and metrics as set to accuracy. Here in fit method epochs is 100 which runs model 100 times to get more precise result as it gets new learned result every time it is run. Prediction having greater than 0.5 is considered a feasible result. By combination of test and predicted result accuracy of result of model is calculated by score_accuracy().

LSTM is a module in TensorFlow library which uses sequential algorithm in Recurrent neural network (RNN) structure which allow use to get more précised output. This result is compared to result from predicted and trained data to get difference in the result percentage accuracy. The design of the organization engineering is represented in Fig.1. In this cycle, the yield is to be utilized as another entrance to finish multi-step estimating. At long last, all yields, prior to going into the circle, are saved as results. The middle is standardized to the shift of zero to 1.

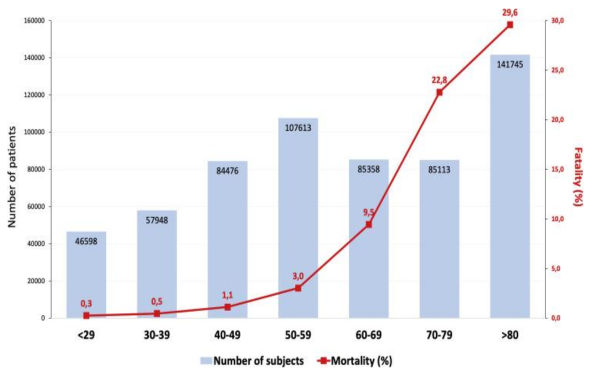
B. Architecture Diagram

Architecture diagram shows different level of steps involved in implementation. According to architecture diagram, Data set is load and process of data visualization is started and missing data processing is done to optimize prediction. Than cleaned data is used to analysis feature of data and data of featured data is extracted. This featured extracted data is engineered and prepared for algorithm. Prediction module uses this data to find behaviour pattern by machine learning algorithm. Hence, predicted output is compared with test data output to find difference in accuracy of module.



V. RESULT

The early expectation of COVID-19 can be useful in decreasing the immense weight on medical care frameworks by assisting with diagnosing COVID-19 patients. In this work, we have used the LSTM with 3 different activation function two of them are Rectified Linear Activation Function and another is signoid for the forecast of COVID-19 disease utilizing the study of disease transmission dataset for positive and negative COVID-19 cases were created. The presentation of model was assessed dependent on exactness boundaries. The model created with LSTM turned out to be the best model among all models created as far as accuracy with 92.94%. The LSTM model showed that the age, Hematocrit, Hemoglobin features are the main element among every one of the dataset's reliant highlights, including the clinical highlights. The model demonstrates that the vast majority over the age of 45 are inclined to be infected with COVID when compared with individuals of lower ages as you can see in the below graph.



VI. CONCLUSION

In this study, the use of Long Short-Term Memory (LSTM) Networks in forecasting the total number of COVID19 cases suits well rather than RNN. We implemented three different types of activation functions for the LSTM Network. The first one was Function (relu) with input_dim 19, the Second was Rectified Linear Activation Function (relu) again without any input_dim and the last one was sigmoid because using the exponential linear unit instead of hyperbolic tangent resulted in better performance. How sigmoid ever, the Addition of column-like level of Hematocrit and level of Basophils of patients resulted in a lower error in predicting compare to using just level of hemoglobin, Platelets, etc as the input of the LSTM Network.

VII. DISCUSSION AND FUTURE WORK

In this, we have just executed ATCS on Spark. This is on the grounds that ATCS is basically structure free, and it applies to any enormous information system that requires boundary tuning. At the point when executed on other huge information structures, like Hadoop, we just need to reselect the boundaries that should be tuned, and afterward rehash the way toward gathering preparing information, preparing the model, and looking for the ideal boundary arrangement. Later on, we will investigate strategies to advance the precision of GAN expectation and plan to consider the relocation of models between various responsibilities. This will additionally decrease the overhead of gathering preparing information and help us assemble a more lightweight programmed tuning framework.

VIII. REFERENCES

[1] L. Wang, S. Tasoulis, T. Roos, and J. Kangasharju, ‘‘Kvasir: Scalable provision of semantically relevant Web content on big data framework,’’ IEEE Trans. Big Data, vol. 2, no. 3, pp. 219–233, Sep. 2016.

[2] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, ‘‘Spark: Cluster computing with working sets,’’ in Proc. HotCloud, 2010,p. 10.

[3] M. Abadi et al., ‘‘TensorFlow: Large-scale machine learning on heteroge-neous distributed systems,’’2016,arXiv:1603.04467. [Online]. Available:http://arxiv.org/abs/1603.04467

[4] D. Sun, G. Zhang, and W. Zheng, ‘‘Big data stream computing: Technolo-gies and instances,’’ J. Softw., vol. 25, no. 4, pp. 839–862, 2014.

[5] Z. Yu, Z. Bei, and X. Qian, ‘‘Datasize-aware high dimensional configu-rations auto-tuning of in-memory cluster computing,’’ in Proc. ASPLOS, 2018, pp. 564–577.

[6] A. Jacob, B. Harris, J. Buhler, R. Chamberlain, and Y. Cho, ‘‘Scalable softcore vector processor for biosequence applications,’’ in Proc. 14th Annu. IEEE Symp. Field-Program. Custom Comput. Mach., Apr. 2006, pp. 295–296.

[7] Z. Fadika and M. Govindaraju, ‘‘DELMA: Dynamically ELastic MapRe-duce framework for CPU-intensive applications,’’ in Proc. 11th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput., May 2011, pp. 454–463.

[8] L. Bao, X. Liu, Z. Xu, and B. Fang, ‘‘AutoConfig: Automatic configuration tuning for distributed message systems,’’ in Proc. 33rd ACM/IEEE Int. Conf. Automat. Softw. Eng. (ASE), Sep. 2018, pp. 29–40.

[9] A. Caprara, M. Monaci, P. Toth, and P. L. Guida, ‘‘A Lagrangian heuristic algorithm for a real-world train timetabling problem,’’ Discrete Appl. Math., vol. 154, no. 5, pp. 738–753, Apr. 2006.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ‘‘Generative adversarial nets,’’ in Proc. NIPS, 2014, pp. 2672–2680.

[11] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, ‘‘A fast and elitist multiobjective genetic algorithm: NSGA-II,’’ IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, Apr. 2002.

[12] Z. Bei, Z. Yu, N. Luo, C. Jiang, C. Xu, and S. Feng, ‘‘Configuring in-memory cluster computing using random forest,’’ Future Gener. ComputSyst., vol. 79, pp. 1–15, Feb. 2018.

[13] N. Yigitbasi, T. L. Willke, G. Liao, and D. Epema, ‘‘Towards machine learning-based auto-tuning of mapreduce,’’ in Proc. MASCOTS, 2013, pp. 11–20.

[14] G. Wang, J. Xu, and B. He, ‘‘A novel method for tuning configuration parameters of spark based on machine learning,’’ in Proc.HPCC/SmartCity/DSS, 2016, pp. 586–593.

[15] Apache Spark. Accessed: Sep. 5, 2019. [Online]. Available:http://spark.apache.org/

[16] M. Zaharia, M. Chowdhury, T. Das, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, ''Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing,'' in Proc. NSDI, 2012,

pp. 15–28.

[17] G. Mackey, S. Sehrish, and J. Wang, ''Improving metadata management for small files in HDFS,'' in Proc. IEEE Int. Conf. Cluster Comput. Workshops, Sep. 2009, pp. 1–4.

[18] Q. Fan, K. Zeitouni, N. Xiong, Q. Wu, S. Camtepe, and Y. Tian, ''Nash equilibrium-based semantic cache in mobile sensor grid database systems,'' IEEE Trans. Syst., Man, Cybern., Syst., vol. 47, no. 9,

pp. 2550–2561, Sep. 2017.

[19] M. Mirza and S. Osindero, ''Conditional generative adversarial nets,'' 2014, arXiv:1411.1784. [Online]. Available: http://arxiv.org/abs/1411. 1784

[20] M. Arjovsky, S. Chintala, and L. Bottou, ''Wasserstein GAN,'' 2017, arXiv:1701.07875. [Online]. Available: http://arxiv.org/abs/1701.07875

[21] C. Hwang and S. Lin, ''Hill climbing for diversity retrieval,'' in Proc. CSIE, 2009, pp. 154–158.

[22] L. Wang, S. Li, F. Tian, and X. Fu, ''A noisy chaotic neural network for solving combinatorial optimization problems: Stochastic chaotic simulated annealing,'' IEEE Trans. Syst. Man, Cybern. B, Cybern., vol. 34, no. 5, pp. 2119–2125, Oct. 2004.

[23] S. Alupoaei and S. Katkoori, ''Ant colony optimization technique for macrocell overlap removal,'' in Proc. 17th Int. Conf. VLSI Design, 2004, pp. 963–968.